ORIGINAL RESEARCH **Open access**

Diversity-in-Experience Training for Deep Reinforcement Learning-Based Building Energy Management

Ahmed Youssef[1*], Khaled Hassan[1], Mahmoud Elamin[2]

Abstract

Deep reinforcement learning (DRL) offers an effective approach to managing uncertainty when supervising the energy consumption of buildings equipped with energy storage devices. Usually, DRL models are trained by selecting samples arbitrarily from a dataset, which frequently leads to an overexposure to specific data classifications alongside a deficiency or complete absence of exposure to other groups. As a consequence, the resulting trained framework can develop biases toward particular data segments, yielding substandard outcomes for categories with which it had minimal interaction. To overcome this limitation, this research introduces an experiential diversity-based training framework for DRL agents. This methodology guarantees that the agents interact with all varieties of information. The proposed system is executed across two distinct stages. In the initial phase, raw data is partitioned into various categories using the K-means clustering method. This clustered data is subsequently structured by layering the information of one group directly over another. In the second phase, a retrieval algorithm is introduced to extract data from each segment to train the DRL agent. The retrieval frequency from each cluster corresponds directly to its total volume of data points, prompting the designation “proportional selection method.” To evaluate the efficacy of the introduced strategy and contrast its performance with traditional arbitrary selection, this study introduces two metrics: the flatness index and the divergence index. The framework undergoes training across multiple distinct data timelines (spanning 1-year, 3-year, and 5-year intervals) while also incorporating solar photovoltaic technology. The simulation results show that the proposed methodology effectively smooths the facility’s power demand curve by optimizing energy storage network usage.

Keywords Deep reinforcement learning, Battery-based power storage, Facility demand management, Diversity of experience, Energy regulation

*Correspondence:
Ahmed Youssef
ayoussef@uofk.edu
[1] Department of Smart Energy Engineering, Faculty of Engineering, University of Khartoum, Khartoum, Sudan
[2] Department of Artificial Intelligence Applications, Faculty of Engineering, Sudan University of Science and Technology, Khartoum, Sudan

Introduction

Buildings represent primary centers of power utilization, accounting for roughly 30% of aggregate energy consumption [1]. The energy requirements of structures grew by nearly 4% in 2021 relative to the preceding year, marking the sharpest surge observed over the prior 10 years [2]. Electricity directly satisfies about 31% of this building energy requirement. As documented in the Electricity Market Report 2023 [3], worldwide demand for electricity expanded by nearly 2% in 2022 compared with the prior year, even amid a global crisis in energy supplies. The same report forecasts that international energy requirements will climb by roughly 3.2% annually starting in 2024, a pace significantly exceeding the pre-pandemic trend of 2.4%. This accelerating growth in power demand is anticipated to stem from the electrification of

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

transport systems [4], the transition of heating networks to electricity [5], and the large-scale industrial manufacturing of substitute fuels like hydrogen [6]. Furthermore, White *et al*. [7] indicate that replacing standard furnaces with reversible heat pumps will increase wintertime electricity needs by as much as 36%.

The immense power usage within buildings transforms them into a primary driver of greenhouse emissions [8], a challenge that can be alleviated through the adoption of renewable energy sources (RES). The implementation of RES, particularly rooftop solar photovoltaic (PV) setups, has recently accelerated [9]. Nonetheless, relying solely on standalone PV generation can cause sharper net load ramp rates [10] because solar power drops off just as residential electricity usage spikes during evening hours. Furthermore, peak solar generation aligns with intervals of lower daytime demand in domestic settings. Thus, incorporating PV systems by themselves cannot sufficiently mitigate peak loads in the evening. To counter this bottleneck, numerous investigations have explored the integration of battery energy storage systems (BESSs), as reviewed below.

An assessment focusing on the financial viability of BESSs within structures [11] determined that operational expenses could be lowered by up to 5.3% through independent BESS deployment. Alternative storage options aimed at regulating the electrical demand of facilities utilizing PV arrays are analyzed in Liu *et al*. [12], which points to lithium-ion batteries as a highly practical option. Diversified structural setups for BESSs, including residential and localized community energy storage systems for building load balancing, are explored in Terlouw *et al*. [13]. Additionally, Koskela *et al*. [14] investigate the economic gains of pairing BESS with PV configurations in structures. The researchers emphasize that a coupled BESS-PV system yields greater financial returns than standalone PV arrays when factoring in all available subsidies. Even so, deploying independent BESS or integrated BESS-PV setups in buildings demands a building energy management system (BEMS) to orchestrate storage operations and lower facility electricity expenses. A variety of research works dedicated to managing structural demand via BEMS are outlined below.

The current body of research on BEMS can be broadly split into two methodologies: model-centric operational strategies [15] and reinforcement learning or deep reinforcement learning (DRL)-based operational strategies [16]. In the first category, an analytical model is formulated to minimize or maximize a designated target function while adhering to a set of operational limits. System unpredictability can be accounted for by using either robust [17] or stochastic optimization frameworks [18]. In the second category, a reward mechanism is established to train a DRL model across varied datasets, and the finalized model is subsequently deployed for real-time adjustments [19].

Various model-centric BEMS setups have been put forward in previous literature. For instance, a framework leveraging Lagrangian multipliers was crafted to maximize financial returns while respecting BESS constraints within a smart grid context in Nge *et al*. [19]. Similarly, a multi-objective optimization routine is designed in Lokeshgupta and Sivasubramani [20] to trim the utility expenses of a domestic building while shaving the peak demand of the broader network. However, the quantity of constraints within model-centric approaches can swell dramatically as more appliances and equipment are introduced. To alleviate this, Ahmad *et al.* [21] propose a strategy to reduce the number of mathematical constraints without sacrificing optimal solution accuracy.

In parallel, multiple investigations have leveraged reinforcement learning and DRL techniques to regulate facility power loads, including setups with BESS and PV assets [22]. For instance, a Q-learning model was engineered in Abedi *et al*. [23] to regulate BESS charging/discharging and trim utility costs for a residential home. Likewise, fuzzy logic principles are embedded within the reward framework of a Q-learning scheme in Alfaverh *et al*. [24] to schedule home appliances in an automated residence, encompassing electric vehicles and BESS configurations. Still, basic reinforcement learning algorithms struggle to handle system variations effectively. Consequently, several investigators have advanced DRL-driven energy management models. For example, deep deterministic policy gradient methods are used in Yu *et al.* [25] and Shen *et al*. [26] to operate BEMS frameworks. The technique detailed in Yu *et al*. [25] operates independently of a dynamic system model, whereas a multi-agent structure utilizing prioritized experience replay is introduced in Shen *et al*. [26].

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Furthermore, the soft actor–critic (SAC) approach is utilized in Pinto *et al*. [27, 28] to orchestrate building loads. Combined energy expenses and peak demand are factored into the model in Pinto *et al*. [27]. In contrast, Pinto *et al.* [28] address average peak demand and the peak-to-average ratio. The utilization of DRL across diverse building categories—such as domestic, corporate, commercial, data storage hubs, and schools—is synthesized in Shaqour and Hagishima [29].

Nonetheless, numerous hurdles must be overcome before RL/DRL systems can be widely implemented for regulating physical buildings. For instance, three core obstacles to deploying RL in physical structures are highlighted in Wang and Hong [30]: the massive volume of training data needed, operational security concerns, and difficulties in applying transfer learning. Similarly, Nagy *et al*. [31] highlight ten primary bottlenecks in deploying RL for building energy regulation in the real world, classifying them into learning, infrastructural, cost-benefit, and safety-security categories. A computer science-oriented analysis on executing RL in buildings is offered in Weinberg *et al.* [32], which categorizes structures into individual and grouped assets while examining single-agent versus multi-agent paradigms. Hurdles in deploying RL for grid-interactive structures are investigated in Nweye *et a*l. [33], where nine major issues are detailed, including poorly formulated reward functions and learning from sparse datasets. The bulk of these studies recognize bottlenecks connected to the productive use of accessible data, such as extensive training data needs and learning from restricted datasets.

As a result, productive exploration remains a tough bottleneck in DRL setups, given that the agent is restricted to learning exclusively from environmental interactions [34, 35]. This issue becomes intensely complex when dealing with expansive state spaces or sparse reward feedback [36]. To remedy this, researchers have incorporated a distance-measure regularization penalty into the loss function to boost the agent's exploratory tendencies [36]. Regional and global objectives are formulated for agents to maximize population variety in Parker-Holder *et al*. [34]. Two distinct generation mechanisms are introduced to amplify the exploration capabilities of DRL agents in Pereira *et al*. [37]. One generator focuses more heavily on exploration while the other prioritizes exploitation, despite sharing an identical structure. In each of these works, the DRL agent's capabilities were enhanced, largely driven by superior exploration.

Importantly, Parker-Holder *et al.* [34], Hong *et al.* [36], and Pereira *et al*. [37] introduce varying methods to alter the internal design of the core architecture. Doing so introduces additional hyperparameters, which can complicate the training phase and require custom fine-tuning for each distinct application. Furthermore, the policy embeddings outlined in Hong *et al*. [36] rely on heuristics and may struggle to adapt to unfamiliar environments, as noted in Parker-Holder *et al.* [34]. Crucially, the approach detailed in Hong *et al*. [36] is customized for gaming domains where feedback is sparse or misleading. Similarly, the frameworks introduced in Parker-Holder *et al.* [34] and Pereira *et al*. [37] are inspired by novelty search techniques that prioritize exploring distinct solution domains. Consequently, these frameworks might favor options that are highly unique but not necessarily practical.

Additionally, none of these methodologies have factored in a well-proportioned blend of training experiences. This specific element is highly critical for building energy regulation, because the demand profiles of structures rarely experience drastic shifts on a day-to-day basis. Hence, a specialized strategy is vital to guarantee diversity throughout the learning phase of DRL agents tailored for managing structural loads.

Furthermore, managing energy within buildings presents a high-dimensional state-space challenge because it requires multi-year data that varies continuously within each operational interval. This complexity arises because facility energy consumption architectures evolve dramatically throughout different seasons of the year. Likewise, power demand patterns fluctuate between working days and holidays, as well as across different hours of the single day. Consequently, a DRL framework must efficiently regulate demand under these highly diverse consumption styles. In contrast, standard practices typically train the DRL agent by picking arbitrary demand profiles out of the dataset [25, 28]. This practice can cause the framework to over-encounter specific load profile sets while failing to adequately encounter or completely ignoring alternative sets, ultimately leading to degraded performance within those neglected groups. Through standard arbitrary sampling, the model might continuously re-examine previously traversed states, thereby restricting its

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

exploration of the aggregate state space [38]. Moreover, this can trigger framework instability owing to rapid variations in the feedback reward. Given that a DRL model acquires knowledge by interfacing directly with its surroundings, exposure to highly varied settings is vital to secure its optimal performance [39]. More precisely, ensuring experiential diversity for the DRL controller across all potential state spaces or environments (which represent load profiles here) is vital. However, establishing this experiential variety is difficult, since the underlying data must be sorted into separate categories (clusters) before being fed into the DRL model. Furthermore, distinct data-extraction routines are necessary during each episode to guarantee that every data classification is represented.

To overcome these hurdles, an experience diversity-driven strategy is introduced for training DRL models tasked with supervising building power demands alongside BESS and solar PV components. Departing from current literature addressing DRL agent diversity [34, 36, 37]—which tends to alter the framework's internal structure—our strategy categorizes the incoming dataset. As a result, the introduced methodology guarantees experiential diversity without modifying the inner network topology of the DRL architecture. Within this framework, the daily electrical profiles of the structure are initially categorized into separate groups. The K-means clustering technique is used to partition these daily load records into K distinct clusters, where the optimal value of K is determined to be 6 in this study. Following this, data points associated with each cluster are isolated and organized by stacking them sequentially. A retrieval logic is then formulated to extract records from each category relative to its total volume (the number of observations it holds), a technique labeled the proportional selection method. These operations can be seamlessly embedded into the data preparation stage. The DRL framework is then trained via this sampled data, guaranteeing that the model encounters various data groups. Two specific metrics, termed the flatness index and the divergence index, are designed to evaluate and contrast the efficacy of this strategy against the standard arbitrary selection approach. The flatness index quantifies how level the load distribution is by contrasting peak demands with valley demands. The divergence index assesses the variance between individual data observations and the daily mean consumption. Computational experiments are performed across diverse timeframes, including 1-year, 3-year, and 5-year data structures, while also accounting for the presence of rooftop PV. Ultimately, comprehensive evaluations are executed for designated days of the year by extracting a single sample from each cluster. For simplicity, the introduced approach utilizing proportional data extraction is abbreviated as the proportional selection method (PSM). In contrast, the traditional arbitrary sampling approach is referred to as the random selection method (RSM) throughout the rest of this text.

Diversity in Experience

The experiential diversity of a DRL model quantifies the degree to which the framework has been trained through exposure to the various available state-space categories or clusters within its environment. Within the DRL paradigm, an agent develops an operational policy to maximize its cumulative reward by interacting with its surroundings [39]. Selecting data items arbitrarily during the training phase can bias the agent's learning toward a specific data type—such as summer working days or winter holidays within the context of this work. To resolve this drawback, data preprocessing is essential to group the information into distinct clusters. Subsequently, a data-extraction routine can sample records from each cluster for training purposes, accounting for the relative volume of each group. Details of the introduced framework are provided in the subsequent section. The following subsections outline the target system layout along with the categorization of data into representative clusters.

System configuration

The setup investigated in this research is a domestic building, which can take the form of an individual detached residence or an apartment unit outfitted with local power storage and optional solar PV capabilities. A structural outline of this system setup is illustrated in **Figure 1**. As shown, the facility integrates a building electrical load, a solar PV array, and a BESS assembly. To maintain simplicity, individual household appliances are not illustrated separately, given that the primary goal of this research centers on training the DRL agent to maximize BESS efficiency. The facility is tied to the utility grid, allowing it to import electricity from the network whenever local energy production falls short of matching the

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

demand. A BEMS takes charge of directing the facility's operations. The DRL framework functions as a core component of this BEMS, and the training sequence for the DRL model under the PSM involves the following phases.

Initially, the building's daily power demand records are compiled over an extended period to capture a wide array of load behaviors across varying seasons and days of the week. This information is then sorted into several distinct groups to ensure each group is fairly represented. Details of this grouping analysis are provided in the upcoming subsection. Next, a retrieval logic is applied to extract training samples from each cluster to feed the DRL agent. Ultimately, a fully trained framework that reflects the insights of all representative data groups is established, ready for optimized building operation with integrated BESS and PV assets.

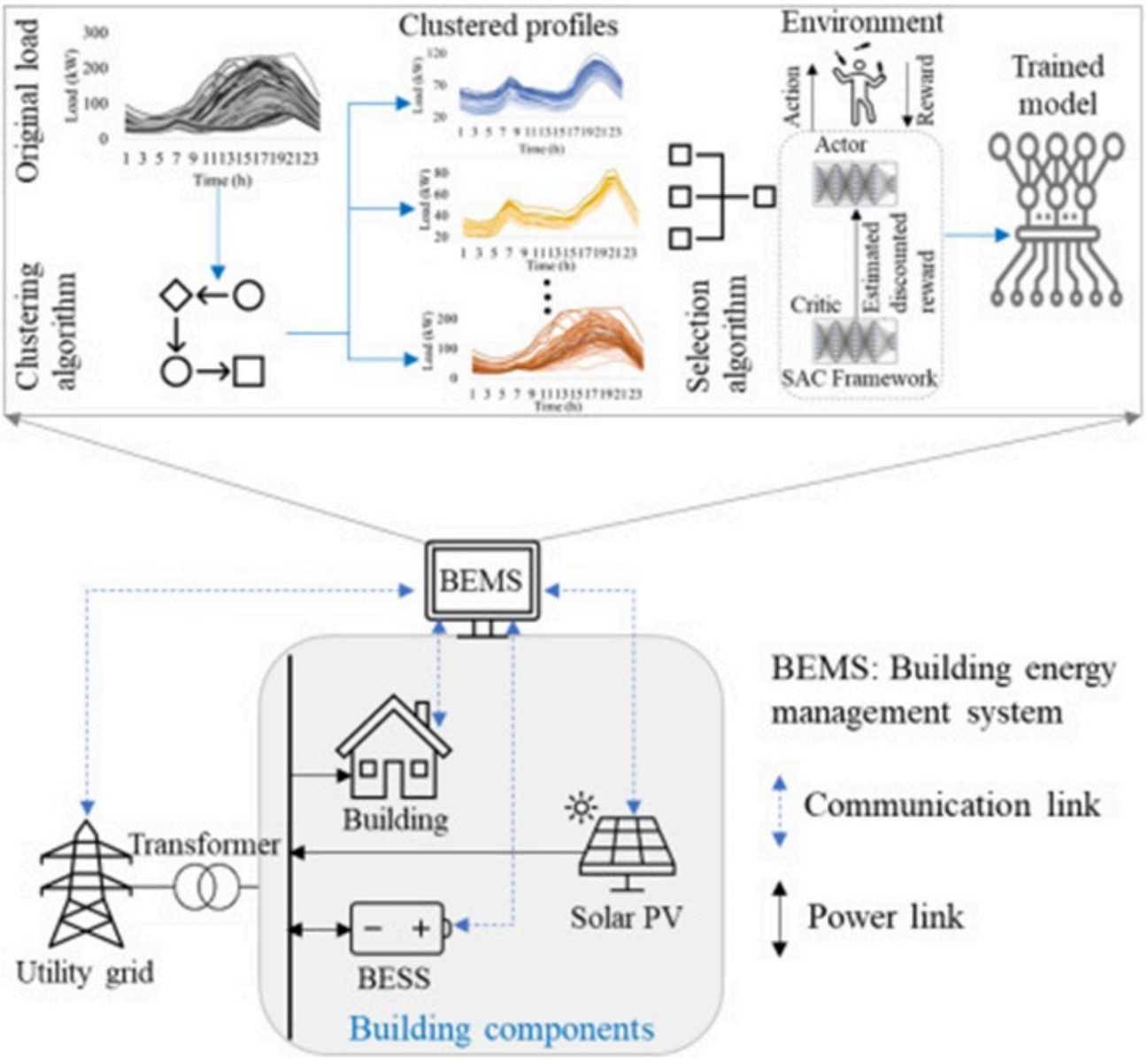


Figure 1. Structural layout of the system and functional blocks of the introduced strategy.

Clustering of daily energy demand profiles

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

The daily power demand records of the facility are initially partitioned into separate groups to facilitate the training of the DRL framework. K-means clustering stands as the most widely implemented unsupervised grouping methodology. Within a K-means framework, an n-dimensional dataset is partitioned into K separate clusters with the primary objective of minimizing the within-cluster sum of squared errors [40]. Typically, the Euclidean distance serves as the metric to evaluate the similarity between two distinct entries [41]. Identifying the ideal number of categories is necessary for any given data repository, after which the total volume of data elements within each cluster can be established. A summary of the steps taken to determine the ideal number of categories for facility load curves using the K-means approach is presented in **Algorithm 1.**

Data: Daily energy consumption profiles of the building for 6 years
Preprocess the data and normalize using the min-max scalar
Initialize the maximum number of clusters (C)
for *j=1:C* **do**
    Randomly determine cluster centroids (c_j)
    while *Centroid position is changed* **do**
        Compute the sum of the squared distance between all data points (1)
        Reassign each element to the cluster to minimize the distance
    end
    Record the distortion for each run
    Record the computation time for each run
end
Result: Optimal number of clusters using the elbow method

Algorithm 1. Optimal number of clusters using the K-means method

Initially, the daily load data of the structure over a designated timeframe (6 years in this investigation) is gathered as input. This data undergoes normalization via a min-max function to prevent higher-magnitude demand curves from dominating the analysis. Following this, the maximum number of testable clusters (C) is established. An outer loop systematically adjusts the number of clusters from 1 to C. For every step j, a cluster center cj is specified, and an inner loop assigns individual data entries to their respective groups. This assignment is carried out by calculating the Euclidean distance (J) using the following formula:

$$J = \sum_{j=1}^{C} \sum_{i=1}^{N} \|x_i^{(j)} - c_j\|^2, \quad (1)$$

Where $x_i^{(j)}$ represents the specific data element i mapped to cluster j, and cj denotes the jth centroid. Each data observation is mapped to its closest center based on the calculated distance J. The cluster centers are subsequently recomputed by averaging all the data points contained within that specific cluster. This iterative procedure continues until the positions of the data points within the clusters stabilize completely. Ultimately, the distortion metrics and processing durations for each configuration are saved, and the ideal quantity of clusters is selected using the elbow technique. The optimal cluster count outcomes are provided in the computational simulation section.

Proportional selection method

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

The daily power curves of the structure are organized into N distinct groups, as outlined in the previous subsection. Following this, a data sampling routine is needed to pull training records from each group throughout the learning cycle. Comprehensive details of this sampling logic are outlined in the next section. Here, the underlying operational concept of the PSM is contrasted against that of the RSM. A conceptual comparison of both sampling strategies is provided in **Figure 2**. For this demonstration, N = 4, and each cluster is distinguished by a unique color. It is likewise assumed that each sequence comprises 4 episodes, where an episode represents the collection of actions and states an agent experiences from its starting point to its concluding state.

Under the traditional approach (RSM), the 4 samples needed for each episode are pulled arbitrarily from the total data pool containing D samples. Consequently, different data clusters tend to dominate across different episodes. For instance, category 1 dominates (comprising 2 out of 4 entries) in sequence 1, whereas category 2 dominates in sequence 2 (comprising 3 out of 4 entries). This traditional approach not only fails to guarantee experiential variety for the model, but it is also highly prone to framework instability due to erratic fluctuations in the average reward within each set.

Conversely, under the PSM strategy, the total number of entries within each cluster is determined first, denoted as D1, D2, D3, etc. Data points are then pulled from each respective group based on the ratio of that specific cluster's volume relative to the entire dataset volume D. For instance, in this scenario, 25% of the samples reside in cluster 1, 50% belong to cluster 2, and the remaining 25% are in cluster 3. As a result, within each training set, two samples are drawn from cluster 2, while one sample is drawn from cluster 1 and another from cluster 3, illustrating the proportional selection method. The PSM guarantees that the model interacts with every data variety by explicitly embedding samples from each category. Furthermore, this approach is likely to stabilize the framework and accelerate the convergence rate of the average reward by maintaining consistent profile distributions across each training cycle.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

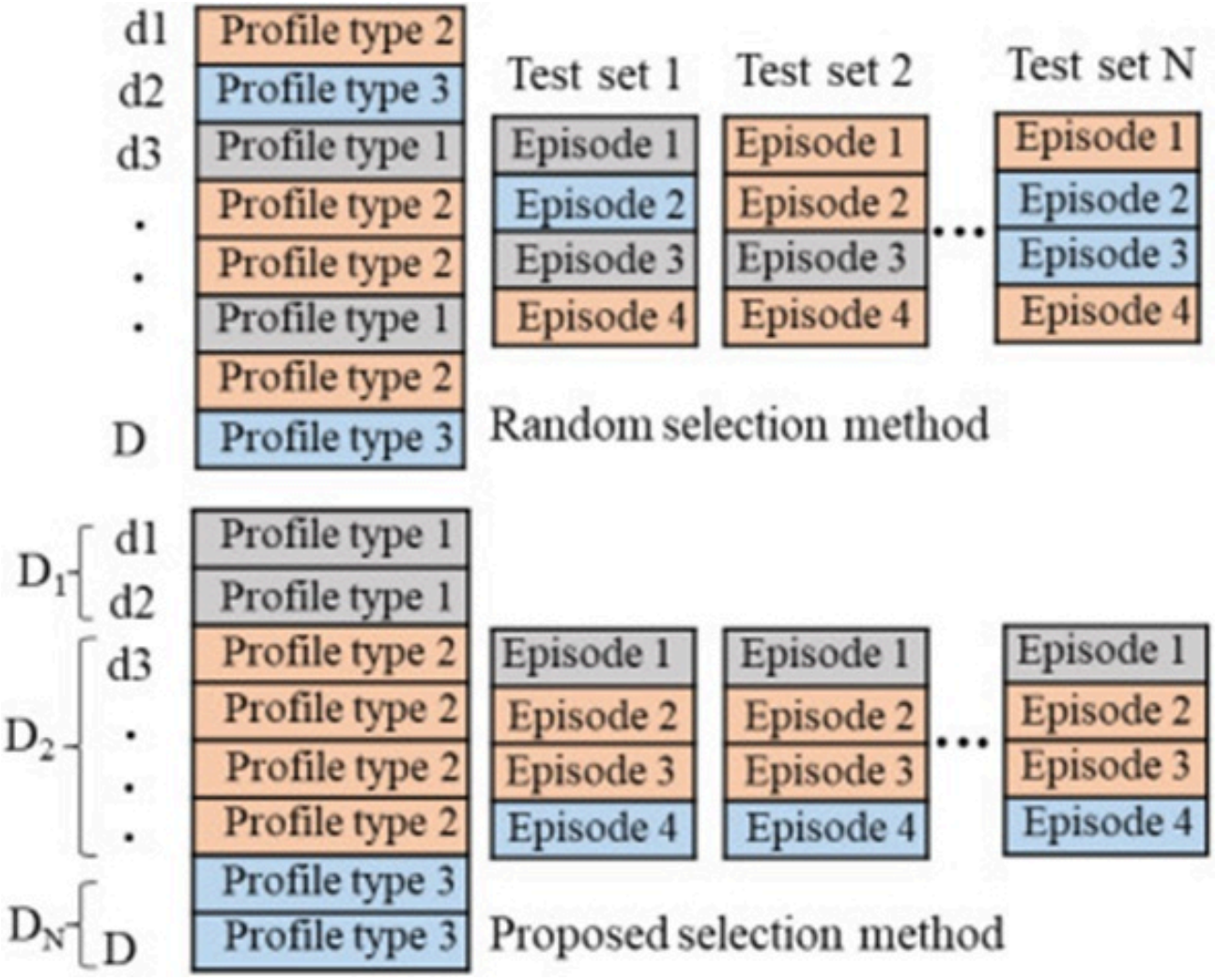


Figure 2. Outline of data sampling methodologies for training the DRL model.

System model

This section builds a mathematical framework for the energy storage network and establishes the governing power balance limits for the electrical system. Note that these expressions cannot be applied directly in their mathematical form, as they are tailored for DRL model training; their functional counterparts will be reconstructed using conditional if-else structures instead. Maximum ($SoC^{\max}$) and minimum ($SoC^{\min}$) state-of-charge boundaries are implemented to prevent overcharging and excessive depletion of the battery unit. Consequently, the power injected into the battery (P_{τ}^{char}) and the power extracted from the battery (P_{τ}^{disc}) during any given time slot t must strictly respect these boundaries. The maximum threshold constraint is enforced through the following formulation:

$$E^{\text{ini}} + \sum_{\tau \le t} \eta^{+} \times P_{\tau}^{\text{char}} \times \Delta t - \frac{P_{\tau}^{\text{disc}} \times \Delta t}{\eta^{-}} \le \frac{B^{\text{cap}} \times SoC^{\max}}{100}, \quad (2)$$

Where E^{ini} represents the starting energy level of the BESS, and B^{cap} denotes the total BESS capacity measured in kWh. The variables η^{+} and η^{-}- signify the charging and discharging efficiencies of the BESS asset, respectively. Lastly, *Δt* represents the time-step duration, in hours, during which the BESS undergoes charging or discharging operations. Similarly, the minimum threshold condition is maintained via the following formulation:

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

$$E^{\text{ini}} + \sum_{\tau \le t} \eta^{+} \times P_{\tau}^{\text{char}} \times \Delta t - \frac{P_{\tau}^{\text{disc}} \times \Delta t}{\eta^{-}} \ge \frac{B^{\text{cap}} \times SoC^{\text{min}}}{100}. \quad (3)$$

The charging and discharging power rates of the BESS are further bounded by the operating capacities of its power converter. These charging (P^{Crate}) and discharging (P^{Drate}) rate limitations are strictly maintained using the following expression:

$$0 \le P_t^{\text{char}} \le P^{\text{Crate}}, 0 \le P_t^{\text{disc}} \le P^{\text{Drate}}. \quad (4)$$

Ultimately, the facility can satisfy its electrical demand (P_t^{load}) by leveraging solar PV generation (P_t^{pv}), drawing power from the main utility network (P_t^{buy}), or discharging the BESS (P_t^{disc}). On the other hand, surplus electricity can be exported back to the utility grid (P_t^{sell}). Finally, the BESS unit can be charged (P_t^{char}) either by importing power from the main grid or by utilizing local solar PV generation.

$$P_t^{\text{buy}} + P_t^{\text{disc}} + P_t^{\text{pv}} = P_t^{\text{sell}} + P_t^{\text{char}} + P_t^{\text{load}}. \quad (5)$$

Soft Actor–Critic-Based Agent Training

In deep reinforcement learning (DRL), a deep neural network is integrated with a traditional reinforcement learning architecture, such as the Q-model. This combined framework is fully capable of managing continuous state spaces as well as environments characterized by high uncertainty, which are challenges that standard Q-learning cannot handle on its own [42].

While multiple DRL variants have been documented across various literatures, the Soft Actor–Critic (SAC) approach stands out as a state-of-the-art framework offering significant benefits over alternative DRL algorithms. These prominent advantages include highly stable operation, accelerated convergence speeds, and strong resilience against getting trapped in local optima [43].

Consequently, this study deploys the SAC algorithm to assess the overall performance of the Profile Selection Method (PSM). The SAC framework serves as the core learning mechanism for evaluating both the traditional baseline approaches and the newly introduced methodologies throughout this research.

Within the SAC framework, the primary objective function is modified by integrating an entropy component designed to evaluate the randomness or predictability of the target random variable. Higher entropy scores denote lower predictability, whereas lower entropy scores indicate higher predictability. A specific tuning hyperparameter \alpha is used to regulate the trade-off between system exploration and exploitation [44].

The learning agent simultaneously develops a policy network (\pi_\theta) alongside a pair of separate Q-networks (Q_1, Q_2). These two action-value networks undergo entirely independent training routines. At the same time, a corresponding pair of target Q-networks is also dynamically updated at each sequential learning interval through a soft-copy update mechanism. The ultimate target of the trained SAC agent is to maximize the combined sum of anticipated future rewards and expected future entropy.

Demand management of buildings with energy storage and solar PV

The comprehensive demand-side management problem for structures equipped with a battery energy storage system (BESS) and solar photovoltaic (PV) generation is handled using the SAC technique due to its specific performance

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

merits, as outlined in the preceding sections. This problem is formally structured as a Markov decision process (MDP), which explicitly defines the states, actions, state transitions, and corresponding reward signals. The environmental state at any given time step t (St) incorporates vital data regarding the active building load (P_t^{load}), the real-time solar PV generation (P_t^{pv}), the current battery state of charge (SoCt), and the specific tracking interval number (t), formulated as follows:

$$S_t = [P_t^{\text{load}}, P_t^{\text{pv}}, SoC_t, t]. \quad (6)$$

To facilitate rapid training convergence and ensure the overarching stability of the neural network learning process, the core state variables are normalized utilizing the minimum ($S^{\min}[k]$) and maximum ($S^{\max}[k]$) bounds of those respective states via the following relationship:

$$S_t[k] = \frac{S_t[k] - S^{\min}[k]}{S^{\max}[k] - S^{\min}[k]}, \quad (7)$$

Where the variable k acts as an operational index used to reference each of the 4 distinct state parameters systematically. Depending on the real-time state of the environment, the learning agent selects a specific control action at time step t (at) restricted within the boundary of [-1, 1]. Inside this range, negative metrics denote a BESS discharging operation, whereas positive values indicate a BESS charging operation. These chosen actions are subsequently converted into the actual target SoC values using the relation below:

$$a_t^{\text{real}} = \frac{a_t}{M} \quad \text{where} \quad M \in Q. \quad (8)$$

Where the real operational choices (a_t^{real}) are calculated by accounting for the specific charging and discharging speed constraints of the BESS power converter unit through the application of a positive rational coefficient M. For instance, setting M = 1 indicates a battery C-rate of 1C, while selecting M = 2 corresponds directly to an adjusted C-rate of 0.5C, following this identical pattern.

The physical state transition advancing from time step t to the next interval t + 1 is dictated by the instantaneous building load demand, the real-time solar PV output, the current SoC level (the active state metric at time t), and the specific management action executed at interval t. This progression can be mathematically formalized as:

$$S_{t+1} = f(S_t, a_t). \quad (9)$$

Ultimately, the localized feedback signal (rt) is computed through a dedicated reward calculation. Extensive details outlining the precise layout of this reward calculation are explicitly detailed in the subsequent subsection.

Reward Function

An all-inclusive reward function (rt) is uniquely formulated to optimize the energy demand profile of a building integrated with both BESS assets and solar PV arrays. The overarching target of the DRL agent is to effectively smooth out the building's net demand curve while carefully avoiding any violations of the core BESS physical energy constraints [45]. The mean daily load demand (P^{avg}) is designated as the primary baseline threshold, and the learning agent receives positive reinforcement for successfully driving the total aggregate load toward this reference baseline. This composite reward function is divided into 3 specific operational segments, as illustrated below:

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

$$r_t = -\frac{(P^{\text{avg}} - (P_t^{\text{net}} + P_t^{\text{bess}}))^2}{P^{\text{avg}}} - p_1 - p_2. \quad (10)$$

The initial component applies a penalty for any structural deviations relative to the reference threshold (the baseline daily average load). This term is normalized utilizing the daily average load metric to scale the resulting daily rewards to an equivalent, uniform range. The net demand parameter (P_t^{net}) highlighted within this primary segment signifies the remaining electrical load of the building after subtracting the locally harvested solar PV power, if available. For a building layout that does not feature solar PV generation, the net load value matches the original, unadjusted load profile exactly. This can be expressed as:

$$P_t^{\text{net}} = \begin{cases} P_t^{\text{load}} - P_t^{\text{PV}}, & \text{if building contains PV,} \\ P_t^{\text{load}}, & \text{otherwise.} \end{cases} \quad (11)$$

Similarly, the variable (P_t^{bess}) represents the actual electrical power charged into or discharged out of the BESS unit at time step t. This metric yields a positive value during active BESS charging cycles and transitions to a negative value during BESS discharging phases. It is mathematically formulated as:

$$P_t^{\text{bess}} = P_t^{\text{char}} - P_t^{\text{disc}}. \quad (12)$$

The final 2 components of the main reward equation represent targeted penalty values designed to punish violations of the pre-established SoC operating bounds. For instance, if the storage unit experiences overcharging (exceeding the defined maximum safety threshold), an explicit penalty (p1) valued at λ_1 is levied, as shown in the following relation:

$$p_1 = \begin{cases} \lambda_1, & \text{if } SoC_t > SoC^{\max}, \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

Identically, an alternate penalty (p2) valued at λ_2 is enforced if the storage asset undergoes a deep-discharge cycle (dropping beneath the lower safe boundary limit). This operational condition is modeled as:

$$p_2 = \begin{cases} \lambda_2, & \text{if } SoC_t < SoC^{\min}, \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

It is important to emphasize that a negative sign precedes every individual component within this overall reward formulation. This structural design implies that all individual terms function strictly as operational penalties; hence, the primary goal of the DRL agent is to minimize these negative values (which mathematically equates to maximizing the total cumulative reward).

BESS Parameters

As established in the preceding section, the operational control action chosen by the agent (at) is converted directly into the target BESS SoC value. The state of charge tracking within the active interval t is continually modified based on the recorded SoC from the immediate past interval (SoC_{t-1}) combined with the deliberate step adjustment executed by the agent, defined as ΔSoC_t. Formally, this dynamic can be written as:

$$SoC_t = SoC_{t-1} + \Delta SoC_t, \Delta SoC_t \in [-1, 1]. \quad (15)$$

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Following this, the total net power transferred into or out of the battery bank (P_t^{bess}) can be derived by analyzing the SoC variance between time steps t and t-1. This relationship is modeled as:

$$P_t^{\text{bess}} = \frac{B^{\text{cap}} \times (SoC_t - SoC_{t-1})}{100}, \quad (16)$$

Where the term B^{cap} represents the total rated energy capacity of the BESS measured in kWh. The specific volume of electrical power directed to charge the BESS throughout the time step t (P_t^{char}) is computed using the following equation:

$$P_t^{\text{char}} = \begin{cases} P_t^{\text{bess}}, & \text{if } P_t^{\text{bess}} > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

Lastly, the following mathematical relation is utilized to quantify the exact volume of power discharged from the battery bank during that same time step t (P_t^{disc}):

$$P_t^{\text{disc}} = \begin{cases} -P_t^{\text{bess}}, & \text{if } P_t^{\text{bess}} < 0, \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

Diversity in experience-based selection and training

This subsection presents the network training architecture in detail. The explicit, step-by-step procedure required to train the DRL agent is systematically itemized in **Algorithm 2**. The primary data streams feeding into this proposed algorithm consist of the building's daily electrical consumption profiles, pre-arranged by cluster groups. First, the ideal number of total clusters (K) is extracted directly via **Algorithm 1.** Next, the total number of individual days in the dataset (D) is computed, and the maximum number of training episodes (N) for agent optimization is defined.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Data: Arranged daily energy consumption profiles of the building: $[D_1, D_2, ..D_i]$
Get the optimal number of clusters (K) from **Algorithm 1**
Get the total number of days in the training set (D)
Set the total number of episodes for training (N)
for i=1:K **do**
    Determine the number of days in cluster i (D_i)
    Determine the day ratio for cluster i (R_i) using (19)
    Compute the cumulative ratio for each cluster i (R_i^c) using (20)
end
for j=1:N **do**
    if MOD$(j, 100) > 0$ AND $\leq R_1^c$ **then**
        Select a random day in the range $[1, D_1]$
    else if MOD $(j,100) > R_1^c$ AND $\leq R_2^c$ **then**
        Select a random day in the range $[D_1, D_2]$
    else if MOD $(j,100) > R_{i-2}^c$ AND $\leq R_{i-1}^c$ **then**
        Select a random day in the range $[D_{i-2}, D_{i-1}]$
    else
        Select a random day in the range $[D_{i-1}, D_i]$
    end
end
Result: Daily energy consumption profiles based on equal experience

Algorithm 2. PSM training process.

The primary execution loop is used to calculate the exact number of days assigned within each cluster i \in K. Subsequently, the proportional day ratio (Ri) matching each cluster group is determined by evaluating its specific day count (Di) against the total aggregated days available across the entire dataset (D = $\sum_{i=1}^{K}$Di). This calculation is mathematically represented as:

$$R_i = \frac{D_i}{\sum_{i=1}^{K} D_i} \times 100. \quad (19)$$

Next, the running cumulative day ratio (R_i^c) for each respective cluster group is derived by calculating the sum of the day ratios for all clusters ranging from index 1 up to the active cluster i. This logic is formulated as:

$$R_i^c = (\sum_{k=1}^{i} R_k). \quad (20)$$

A secondary execution loop then iterates continuously across the total designated number of training episodes (N). For each episode j, a target load profile sample must be drawn from one of the established cluster groups. The selection of this specific profile sample from any given cluster is governed entirely by the value of the cumulative ratio computed through **Eq. 20**. To illustrate, the first R_1^c set of profile samples will be drawn exclusively from cluster 1, R_2^c the

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

subsequent samples from cluster 2, and continuing systematically. This sampling strategy resets every 100 episodes and repeats continuously until the entire training episode allocation is exhausted.

Indices for performance analysis

To properly evaluate the overall efficacy of this profile selection methodology, this research utilizes 2 distinct performance metrics. These identical evaluation criteria are also applied to rigorously benchmark the performance of the PSM against the standard Random Selection Method (RSM). The primary metric is designated as the flatness index (FI), which explicitly measures how uniform the electrical demand curve remains over time. It can be mathematically modeled by using the peak daily load value (P^{max}) alongside the minimum daily load value (P^{min}):

$$\mathrm{FI} = 100 - \left(\frac{P^{\max} - P^{\min}}{P^{\max}} \times 100\right). \quad (21)$$

This metric represents a continuous scale that spans between [0, 100] when analyzing facilities that do not possess on-site solar PV arrays. Higher resulting values correspond directly to a smoother, flatter energy demand curve, whereas lower values indicate the inverse. It is worth noting that the calculated value of FI can drop below zero for building configurations that integrate solar PV systems.

The secondary evaluation metric is called the divergence index (DI), which quantifies the structural deviation of the net load profile (P_t^{net}) relative to the baseline daily average demand (P^{avg}). It is mathematically written as follows:

$$\mathrm{DI} = \frac{\sqrt{\sum_{t=1}^{T}(P_t^{\mathrm{net}} - P^{\mathrm{avg}})^2}}{T} \quad (22)$$

Similar to the primary index, this metric is also configured as a continuous calculation, where elevated values indicate greater structural divergence from the baseline, and minimized values represent closer convergence. Consequently, achieving a high system performance relies on maximizing the FI scores while minimizing the DI scores.

Performance Evaluation

This section details the preprocessing of the raw data and its organization into distinct groups using the clustering techniques described earlier. Following this, the performance of the Profile Selection Method (PSM) is evaluated against the Random Selection Method (RSM) during the training phase of the deep reinforcement learning (DRL) agent. Both the proportional framework and the standard DRL configurations are coded and executed using Python 3.9.13.

Input data

The foundational dataset contains actual power consumption measurements collected from a residential housing complex situated in the USA [46]. The tracking information utilized throughout this work features an hourly (1 h) sampling interval spanning 6 consecutive years, representing a scaled variant of the data originally presented in Taheri *et al*. [46]. Corresponding solar photovoltaic (PV) generation profiles across the identical multi-year timeframe are modeled and produced via Renewables Ninja [47].

An illustrative overview of this unrefined dataset, highlighting the variation among different load profile groups, is depicted in **Figure 3.** To provide clear visualization, specific load and PV power curves tracking various days and distinct seasons are explicitly illustrated in **Figure 4**. As expected, the solar PV generation shows increased energy yields (characterized

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

by longer daylight durations and higher generation peaks) throughout summer cycles, while showing lower energy outputs (characterized by shorter daylight periods and lower peaks) across winter periods.

Similarly, the building demand configurations show substantial variations depending on the season. Furthermore, discrepancies between typical weekdays and holidays are clearly visible within identical seasonal blocks. This variation implies that the operation of the battery energy storage system (BESS) must be dynamically adjusted to accommodate shifting seasons and different days of the week.

A BESS capacity of 250 kWh is chosen by factoring in both the building demand profiles and the solar PV generation characteristics. The operating thresholds for the battery state of charge are established with an upper boundary of 90% and a lower boundary of 10%. Ultimately, the continuous charging and discharging rates of this BESS asset are constrained to C/5, indicating that the storage system can transition between a completely depleted and fully charged state within a 5 h period.

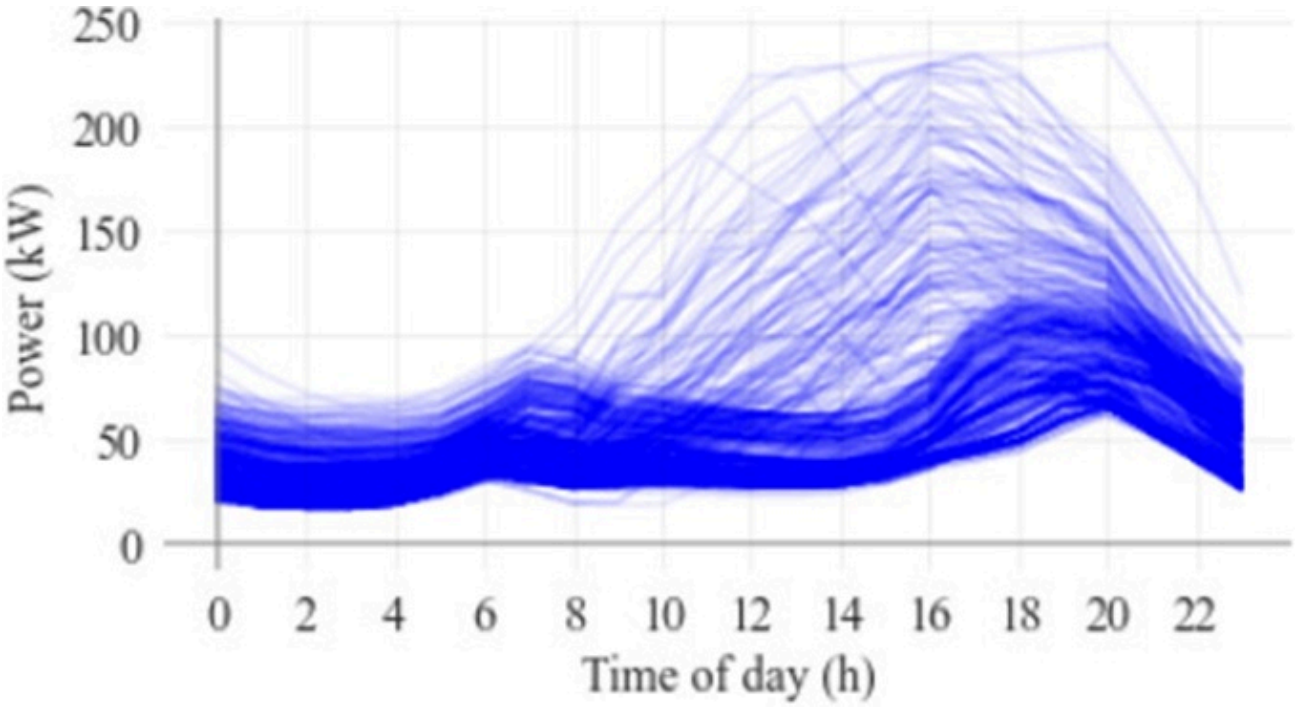


Figure 3. Overview of raw data used in this study.

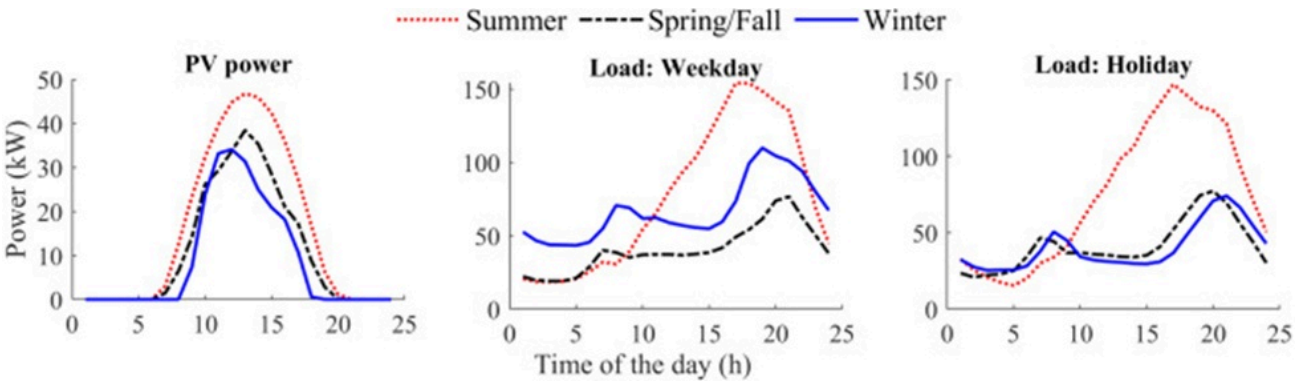


Figure 4. Daily $_{load}$ profiles of selected days for different day types and seasons.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Analysis of optimal number of clusters

In this subsection, **Algorithm 1** is deployed to determine the ideal number of clusters needed to categorize daily energy consumption patterns into discrete sets. The target cluster count is incrementally varied from 1 to 12, with the resulting distortion levels and computational processing durations (fit time) logged for every separate execution. The compiled parameters are mapped out in **Figure 5**, revealing that the recorded distortion values consistently decrease as the total number of allowed clusters expands.

Because the distortion index quantifies the statistical variance separating an individual data sample from its assigned cluster centroid, it is logical that incorporating a larger volume of centroids (clusters) minimizes this overall distortion. Nonetheless, the decline rate of the distortion metric is non-linear; it initially drops rapidly before entering a flatter trajectory.

The elbow optimization framework is implemented to locate the specific inflection point where this structural change takes place to determine the most effective number of clusters [48]; for this dataset, this occurs at K = 6. Matching the analysis found in Teoh and Rong [49], the necessary fit time is also plotted within **Figure 5.** This processing metric reveals an initial drop, after which it scales upward in tandem with an expanding cluster count.

[Elbow Optimization Trend Summary]

Low Clusters (1-5) --> High Distortion, Fast/Variable Fit Time

Optimal Point (K=6) --> Inflection Point (Elbow Location)

High Clusters (7-12) --> Flattened Distortion, Rising Computational Fit Time

Once this ideal cluster threshold is pinpointed, the unrefined input profiles **(Figure 3)** are sorted into separate cluster groups. Every single cluster is visually distinguished via a unique color code, as illustrated in **Figure 6,** where the dashed lines tracking each color pattern mark the mathematical centroids of those respective clusters.

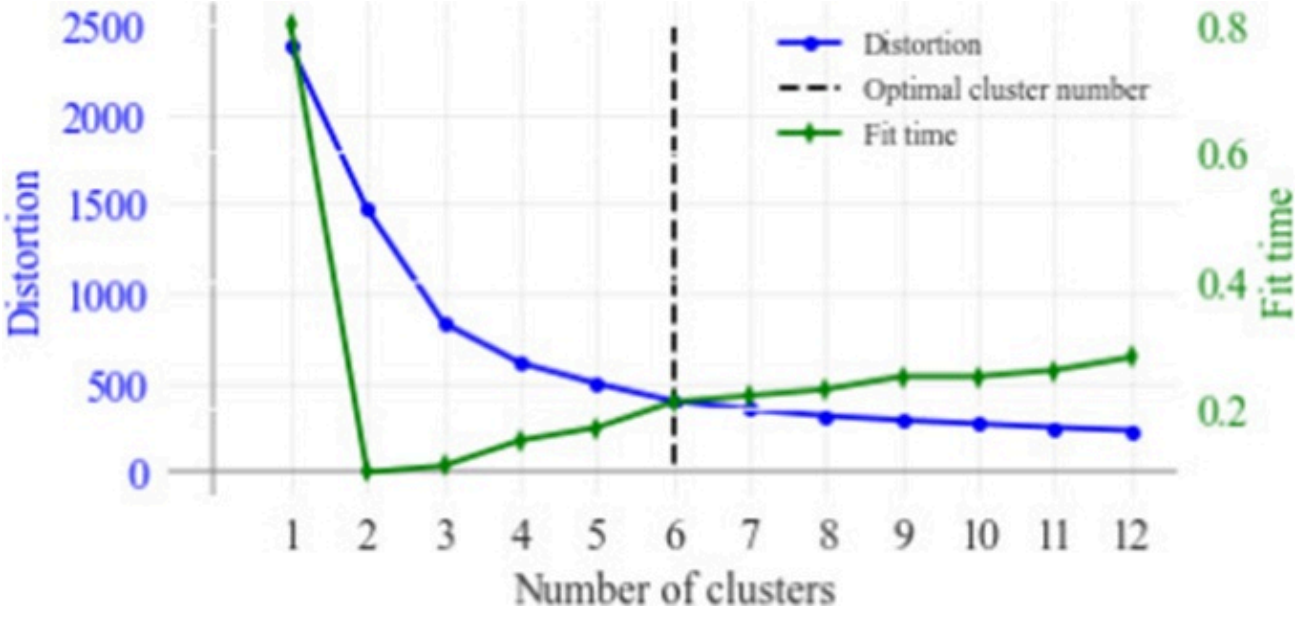


Figure 5. Determining the optimal number of clusters using the Elbow method.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Generally, advanced dimensionality reduction frameworks are implemented to confirm and validate the precision of clustering outputs. The t-distributed stochastic neighbor embedding (**t-SNE**) protocol [50] represents an effective and frequently applied method for this purpose [51]. Consequently, this technique is integrated into this research to reduce the data dimensionality from 24 dimensions to 2 dimensions.

The resulting visual mappings are detailed in **Figure 6**, where every isolated dot represents a single daily load profile. The color assignments mirror the exact cluster groups displayed in **Figure 7.** The close physical grouping of identically colored data points proves that the K-means approach successfully segregated the dataset into distinct clusters.

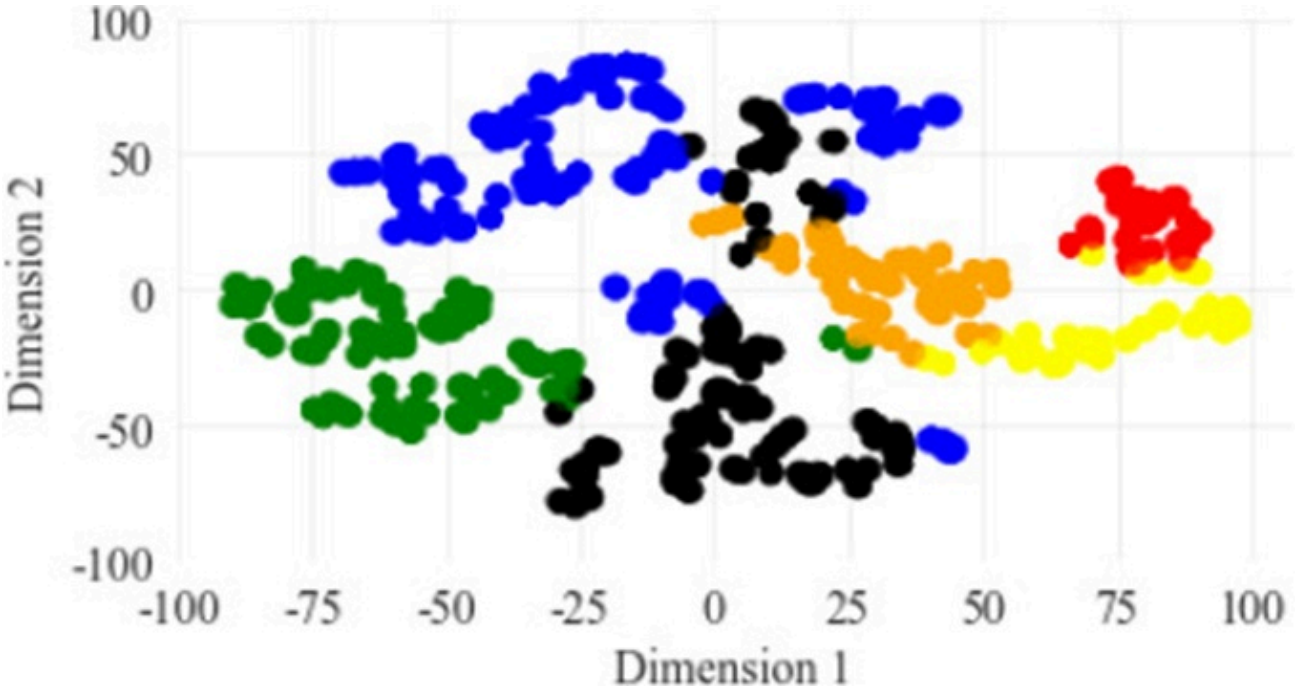


Figure 6. Overview of results obtained from the t-SNE method with two dimensions.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

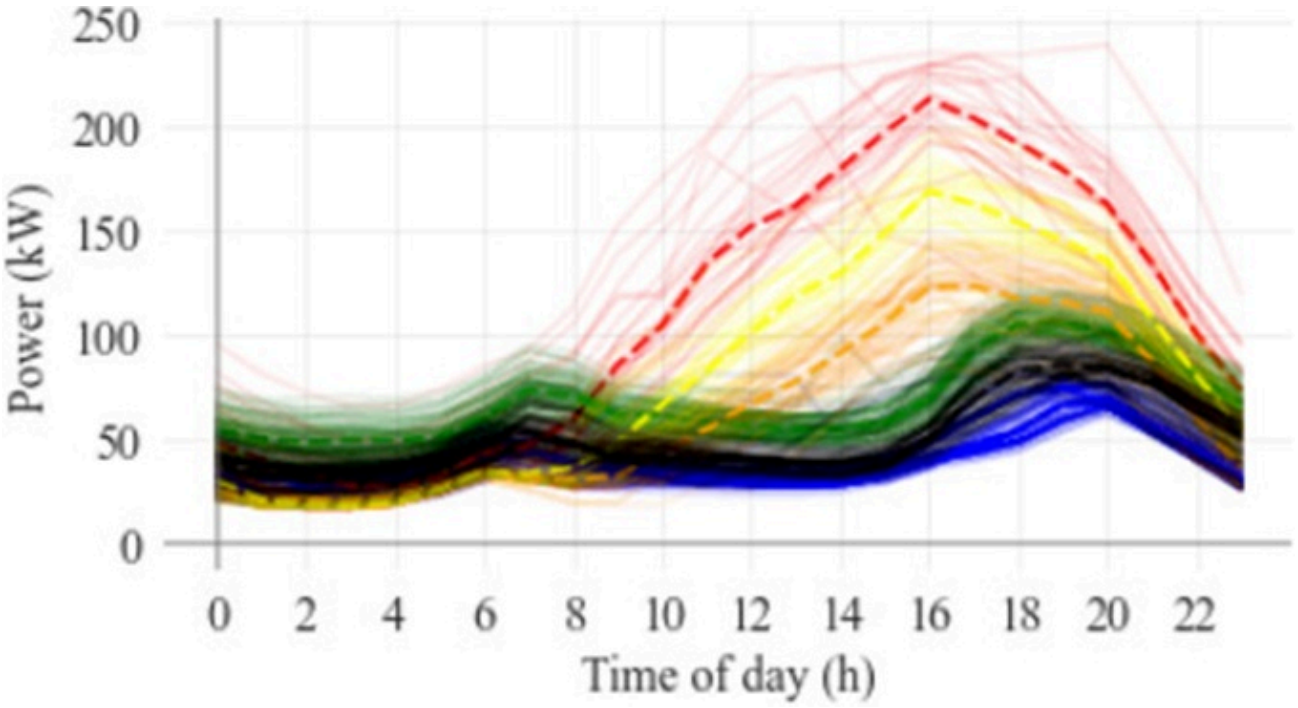


Figure 7. Grouping of load profiles into different clusters.

A comprehensive breakdown detailing the total data points contained within each independent cluster (Di), the proportional day ratio matching each cluster (Ri), and the running cumulative day ratio for each group (R_i^c) is compiled inside **Table 1**. Additionally, the dominant calendar timeframes across the year (encompassing both seasonal weather patterns and specific day classifications) for every cluster group are itemized within this same table.

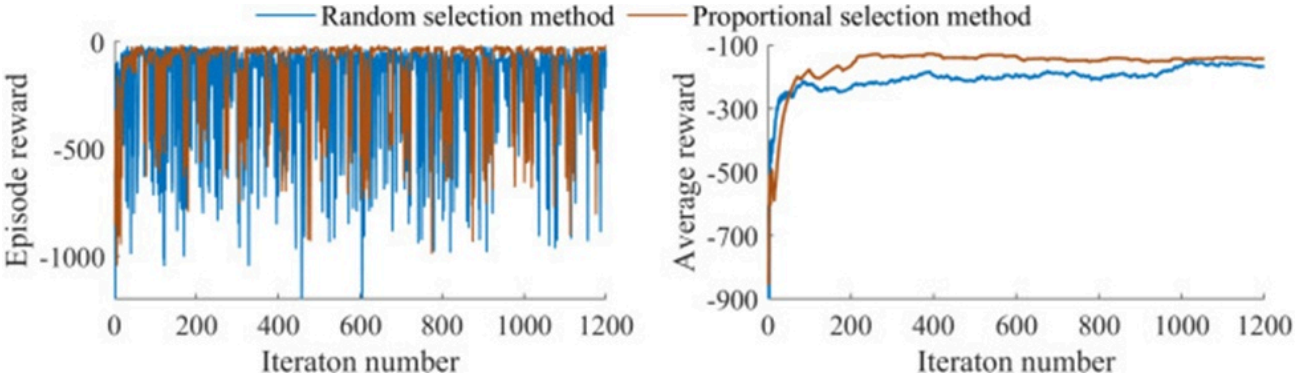
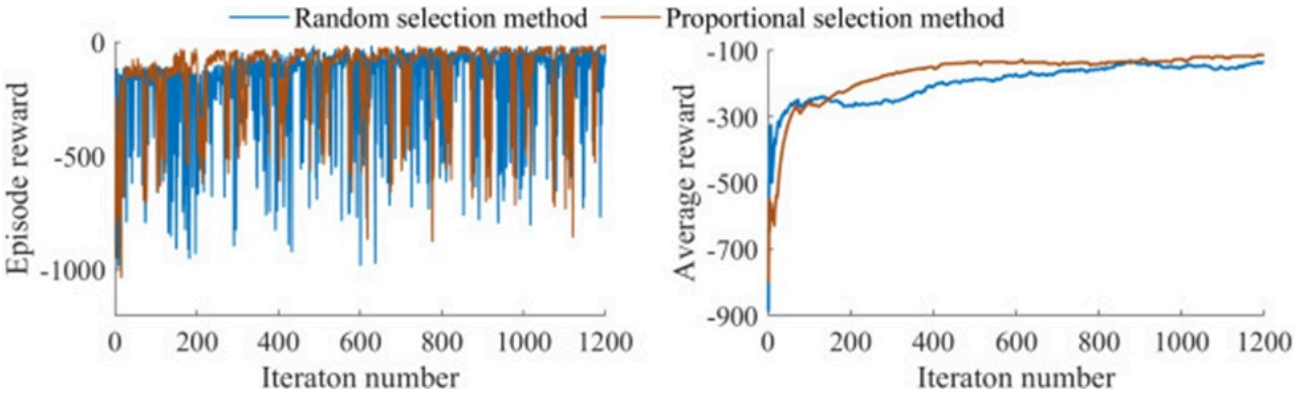


Figure 8. Convergence and stability results with 1-year data.



© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Figure 9. Convergence and stability results with 3-year data.

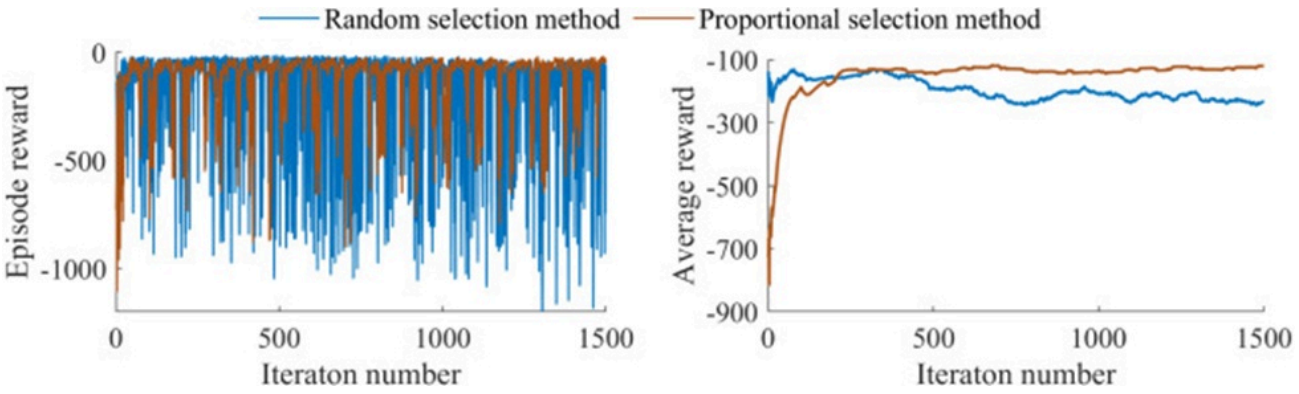


Figure 10. Convergence and stability results with 5-year data.

Table 1. Summary of data samples and related statistics in each cluster.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Time period	**Weekdays**			**Holidays**		
Empty Cell	**Summer**	**Spring/Fall**	**Winter**	**Summer**	**Spring/Fall**	**Winter**
D_i	510	370	625	214	164	309
R_i (%)	23	17	29	10	8	14
(%)	23	40	69	78	86	100

Convergence and stability analysis

The primary DRL agent is optimized utilizing an aggregated dataset tracking a 5-year duration (as outlined in the previous subsection). Independent testing data from the remaining 6th year is reserved exclusively to verify operational performance, as explained in the subsequent section. To analyze the efficacy of the proposed selection methodology, the underlying architecture is trained across multiple dataset scales, specifically featuring 1-year, 3-year, and 5-year data structures without incorporating solar PV generation (focusing strictly on the BESS asset). Ultimately, a final scenario is simulated that simultaneously integrates both the BESS asset and solar PV arrays within the building framework.

The convergence and stability records collected from the first three operational scenarios (featuring the 1-year, 3-year, and 5-year datasets) are presented in **Figures 8-10**, respectively. The left-hand panels of these charts display the raw rewards earned per episode, whereas the right-hand panels plot the running average rewards calculated across a 100-episode window.

It is evident that the PSM configuration systematically outperforms the RSM baseline across all three test scenarios. The operational superiority of the PSM approach over the RSM strategy is demonstrated through higher average reward accumulation, superior system stability, and faster training convergence speeds. The PSM framework converges on higher reward levels in all evaluated scenarios, demonstrating that it has thoroughly mapped the target environment and identified superior operational actions (specifically regarding BESS charging/discharging rates and execution intervals).

Furthermore, its tracked mean reward exhibits significantly smaller variations compared to the traditional baseline method, demonstrating a more stable learning trajectory. Finally, the PSM configuration reaches convergence faster in every scenario. Notably, the RSM baseline shows a clear downward trend as the volume of training data points expands, as seen in the 5-year simulation case.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

The performance profiles for the facility integrating both BESS and solar PV assets are mapped out in **Figure 11**, where the underlying agent undergoes training across the 3-year data allocation. An identical performance trend surfaces here, matching the behavior observed across the prior three test configurations. The PSM framework beats the RSM framework across all categories, including training stability, convergence acceleration, and overall reward metrics. This evaluation validates that the proposed selection strategy maintains superior adaptation across varying environment profiles relative to the RSM baseline. The distinct control actions that yield these varying reward outputs are evaluated and contrasted within the following sections.

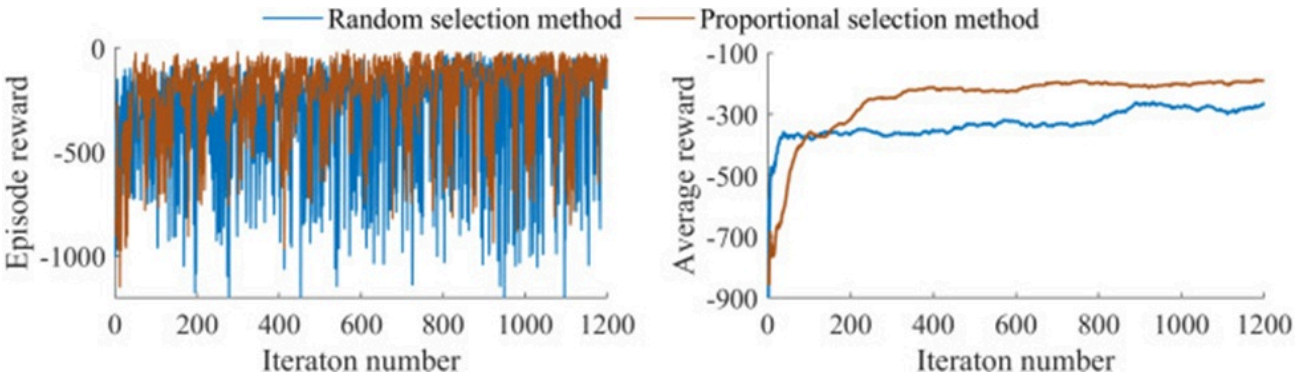


Figure 11. Convergence and stability results with inclusion of PV.

Discussion and Analysis

This section compares the performance of the PSM against the RSM using the two evaluation indices. Diverse operational scenarios are examined to analyze the performance of the PSM, including varying dataset sizes, the integration of solar PV generation, and isolated daily profile evaluations.

Analysis with different datasets

In this subsection, three distinct datasets are assessed to analyze the performance of both the proposed PSM and the standard RSM using the data shown in **Figure 4**. The metrics detailing the flatness and divergence indices derived from the 1-year data allocation are plotted in **Figure 12.** In an identical manner, **Figures 13 and 14** map out the corresponding evaluations using the 3-year and 5-year datasets, respectively.

As expected, the calculated flatness index is lowest for the unadjusted baseline case across all simulated scenarios. This unadjusted case represents the raw electrical demand before integrating any BESS support. The recorded flatness index shows an improvement for the RSM configuration relative to the unadjusted baseline due to the physical addition of the BESS asset.

However, across all test cases, the PSM setup achieves the highest overall flatness index. Elevated flatness index scores denote a much more uniform energy demand curve, which represents the primary target. The most significant variance observed in the flatness index reached approximately 32% when measured against the RSM (within cluster 2) and roughly 63% when evaluated against the unadjusted baseline (within cluster 4).

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

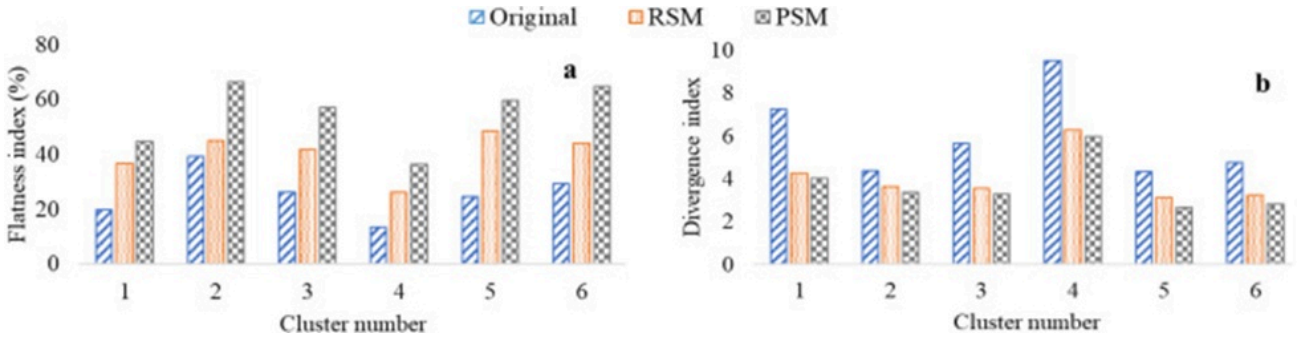


Figure 12. Load variation results with 1-year data: (a) flatness index, and (b) divergence index.

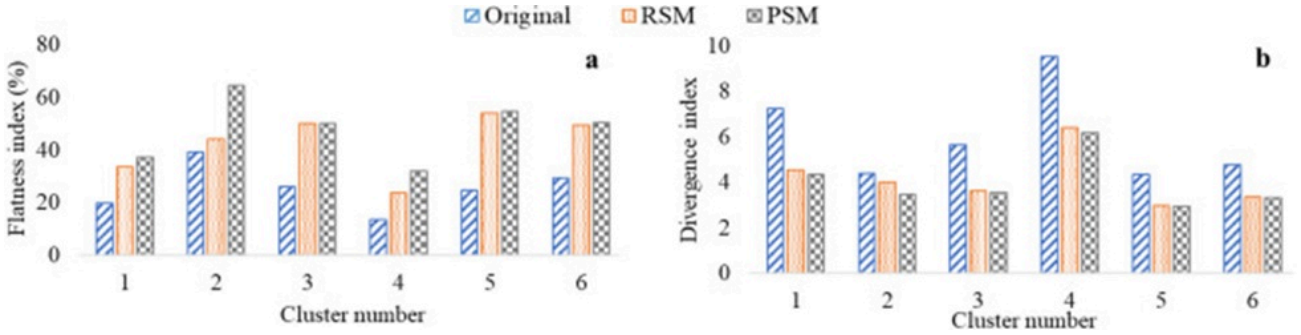


Figure 13. Load variation results with 3-year data: (a) flatness index, and (b) divergence index.

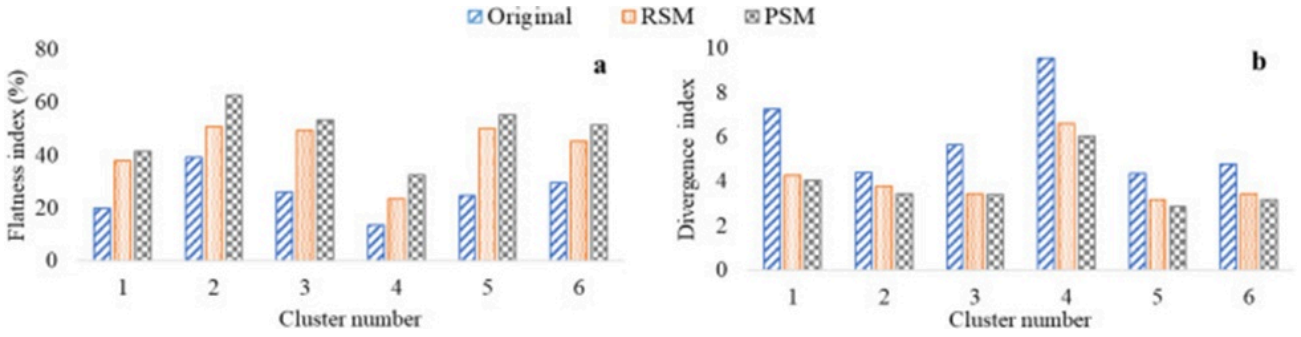


Figure 14. Load variation results with 5-year data: (a) flatness index, and (b) divergence index.

The corresponding divergence index evaluation reveals that the PSM configuration achieves the lowest values across all scenarios. Because higher divergence index (DI) values indicate wider deviations from the baseline, lower scores are preferred. The largest observed difference reached close to -15% relative to the RSM baseline (within cluster 5) and around -45% compared to the unadjusted baseline scenario (within cluster 1).

This assessment confirms that the PSM framework delivers superior performance compared to the traditional RSM across both the flatness and divergence metrics. This outcome matches the convergence analysis, where the PSM consistently stabilizes at higher reward levels, corresponding to better utilization of the BESS capacity to smooth out the building's demand curve.

Analysis with integration of PV

This section compares the performance of the PSM against the conventional RSM while accounting for the integration of solar PV arrays. Mirroring the previous section, the flatness and divergence metrics are analyzed. As documented in **Table 2**, the PSM yields the highest flatness index across all test cases, achieving the desired target.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

Similarly, the divergence metrics detailed in **Table 3** show that the proposed approach secures the lowest divergence index scores across all evaluated scenarios. Minimizing the DI metric is a priority, as it indicates a smaller deviation from the average daily demand. This analysis demonstrates that the proposed selection strategy maintains superior performance even when solar PV generation is integrated into the building's energy ecosystem.

Table 2. Flatness index results (%) with the integration of BESS and PV.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Original	19.7	32.5	39.9	15.7	40.2	41.6
With PV	−6.3	14.2	19.0	13.9	9.4	11.3
RSM	17.4	32.4	44.4	12.4	45.8	43.7
PSM	29.9	64.2	57.9	20.9	55.0	50.7

Table 3. Divergence index results with the integration of BESS and PV.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Original	7.3	4.1	4.5	8.2	4.0	5.2
With PV	8.5	4.9	5.3	8.8	5.4	6.7
RSM	5.5	3.5	3.9	7.5	3.8	4.4
PSM	5.1	2.9	3.4	6.3	3.5	4.2

Profile comparisons

This subsection examines the daily energy consumption profiles for selected days of the year, choosing one representative day from each cluster group. For clear visualization, a single scenario featuring the 3-year dataset is extracted from the various data configurations and plotted in **Figures 15 and 16**. Additionally, the operational profiles detailing the solar PV integrated scenario are evaluated in **Figures 17 and 18**. All outcomes illustrated in this section are generated using the fully trained SAC agent.

Every discrete environment state incorporates comprehensive information tracking the active building load (P_t^{load}), the real-time solar PV output (P_t^{pv}), the battery state of charge (SoC_t), and the current interval number (t). In the same manner, the control action chosen by the agent (at) is transformed into a target battery SoC via **Eq. 8**. This derived SoC value is subsequently applied to compute the exact volume of power charged into or discharged from the BESS during that specific interval t using **Eq. 16**. Furthermore, the agent is governed by penalty metrics designed to punish any violations of the pre-established SoC operating bounds.

To generate the results discussed in this subsection, the trained agent is supplied with load and solar PV generation data tracking a continuous 24-hour window (one full day) for each cluster. The agent's interval-specific SoC profile (action) is logged and applied to determine the resulting battery power adjustments. This procedural sequence yields the profiles shown in **Figures 16 and 18.**

In a matching step, adjusted net load profiles are calculated by combining the raw building load (P_t^{load}), the solar PV generation (P_t^{pv}), and the responsive battery power output (P_t^{bess}), structured as P_t^{load} - P_t^{pv} +P_t^{bess} . The modified demand curves matching both the PSM and RSM frameworks, illustrated in **Figures 15 and 17**, are derived through this

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

approach. It should be noted that P_t^{bess} yields a positive value during active charging cycles and switches to a negative value during discharging operations. Consequently, the net electrical demand scales upward during charging intervals and decreases during discharging intervals.

\text{Adjusted Net Load} = P_t^{load} - P_t^{pv} + P_t^{bess}

- Charging interval: $\mathrm{P}_{\mathrm{t}}^{\mathrm{bess}}$ > 0 \rightarrow Increases Net Demand
- Discharging interval: $\mathrm{P}_{\mathrm{t}}^{\mathrm{bess}}$ < 0 \rightarrow Decreases Net Demand

The daily load shapes tracking the unadjusted baseline, the standard RSM, and the proposed PSM are illustrated in **Figure 15**. The matching charging/discharging power levels and SoC trajectories of the BESS are plotted in **Figure 16**. Note that positive values in **Figure 16** represent charging power, while negative values represent discharging power. The enhanced smoothness of the adjusted load profiles relative to the unadjusted baseline curve is apparent.

The variation separating the charging and discharging strategies of the RSM and PSM configurations is visible in **Figure 16**, which directly causes the discrepancies observed between the final load shapes in **Figure 15**. The proposed selection approach reaches its maximum SoC ceiling later across all cases compared to the RSM baseline. This demonstrates that the BESS charges across a wider interval window using reduced power steps per interval, resulting in a flatter net load profile.

It can also be observed that the peak-to-valley load variance remains lower for the PSM setup compared to the RSM baseline in all instances. The term “peak” characterizes the absolute maximum value occurring within the load curve, while “valley” marks the absolute minimum value. Consequently, the “peak-to-valley” metric defines the total structural gap separating the highest and lowest demand points.

Minimizing this peak-to-valley gap is preferred, as it signifies a more balanced, uniform distribution of electrical demand. The scores obtained across various evaluation criteria (such as the flatness index and divergence index) validate these graphical trends.

The modified load shapes tracking the solar PV integrated scenario are illustrated in **Figure 17,** with the corresponding BESS charging/discharging power curves and SoC tracking mapped out in **Figure 18**. The resulting net load profiles are significantly flatter following the integration of the BESS asset compared to both the unadjusted baseline and the standalone solar PV configuration. Notably, the net load curve (accounting for solar PV) drops below zero within Cluster 1.

Additionally, the standalone solar PV profile introduces a wider peak-to-valley discrepancy relative to the raw, unadjusted building load. The proposed selection approach yields smoother demand curves compared to the RSM framework. It is also evident that the RSM baseline fails to fully leverage the available BESS capacity across most scenarios, failing to charge the battery up to its maximum SoC limit.

In contrast, the proposed PSM fully utilizes the BESS capacity in every test case. This limitation explains the lower performance of the RSM baseline when evaluated across the various performance indices detailed in the previous section.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

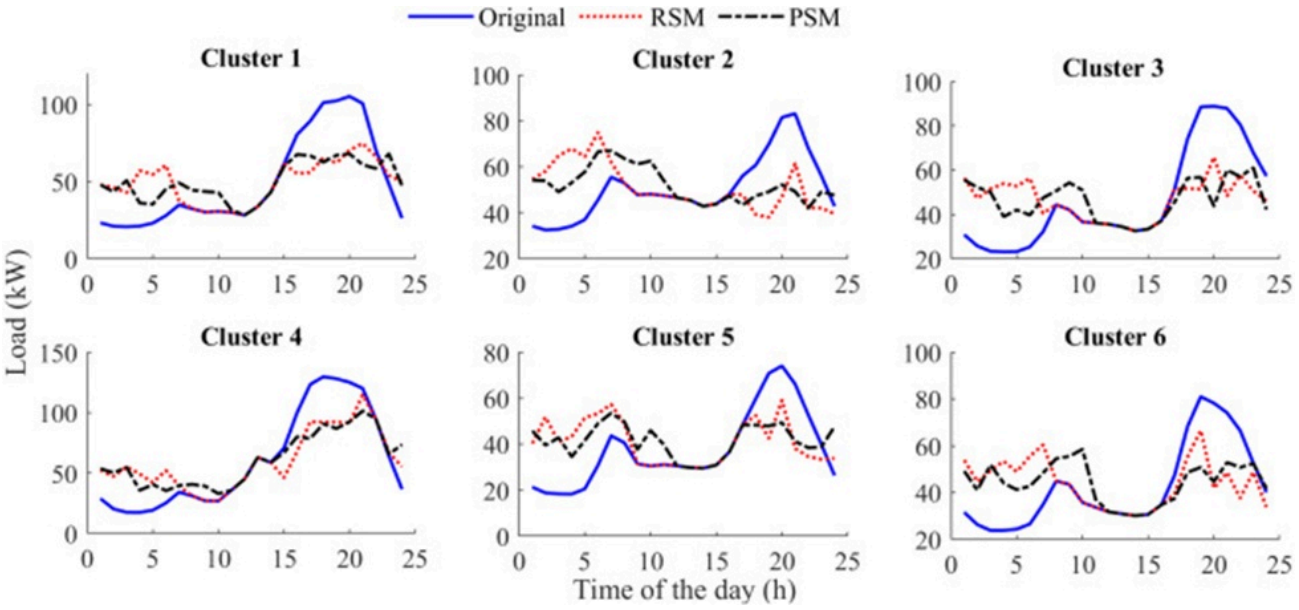


Figure 15. Load profiles of selected days with model training using 3 years of data.

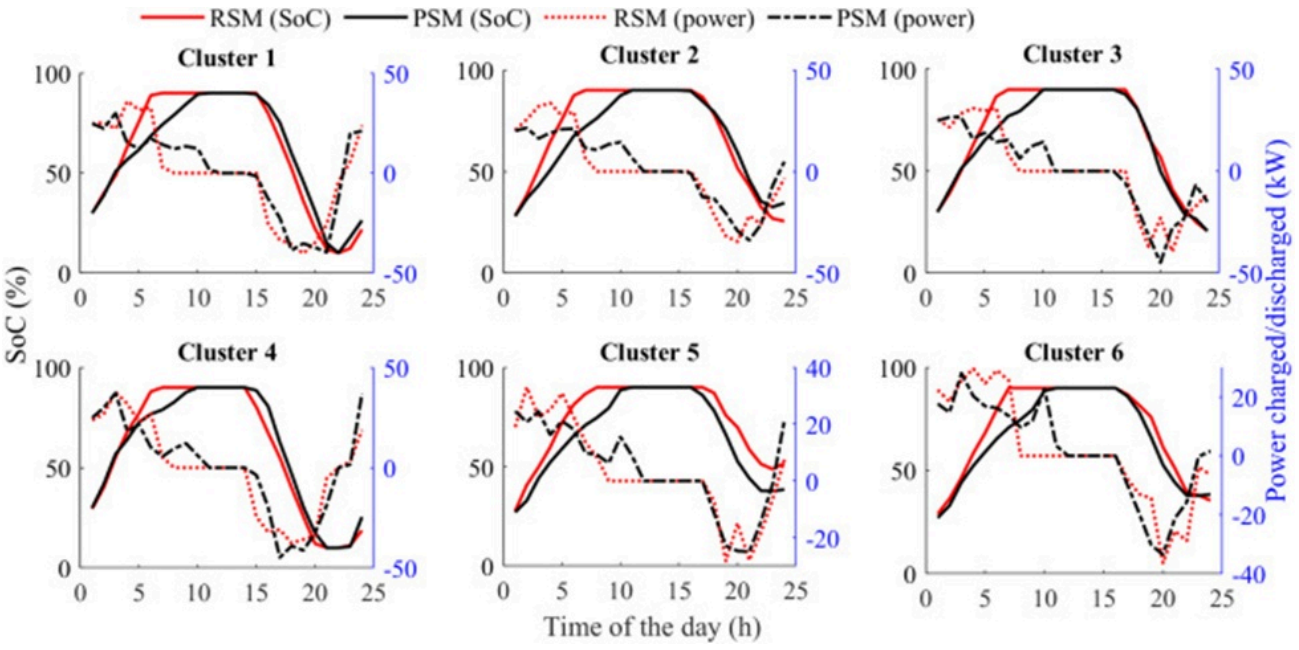


Figure 16. BESS power and SoC profiles with model training using 3 years of data.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

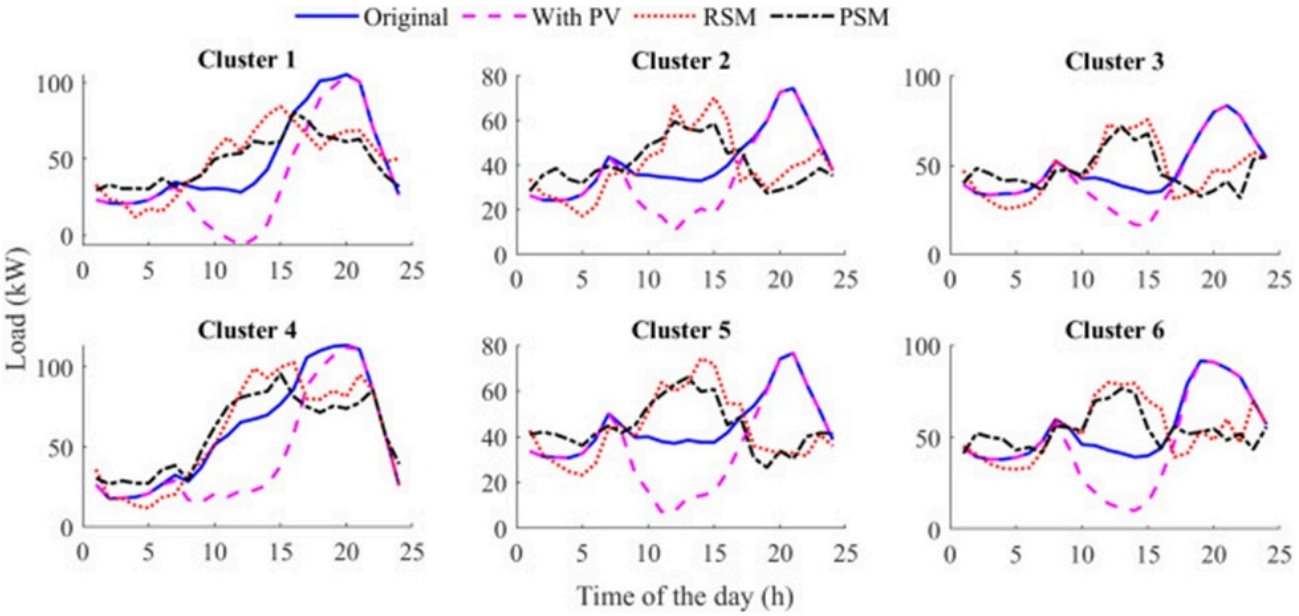


Figure 17. Load profiles of selected days with the integration of PV.

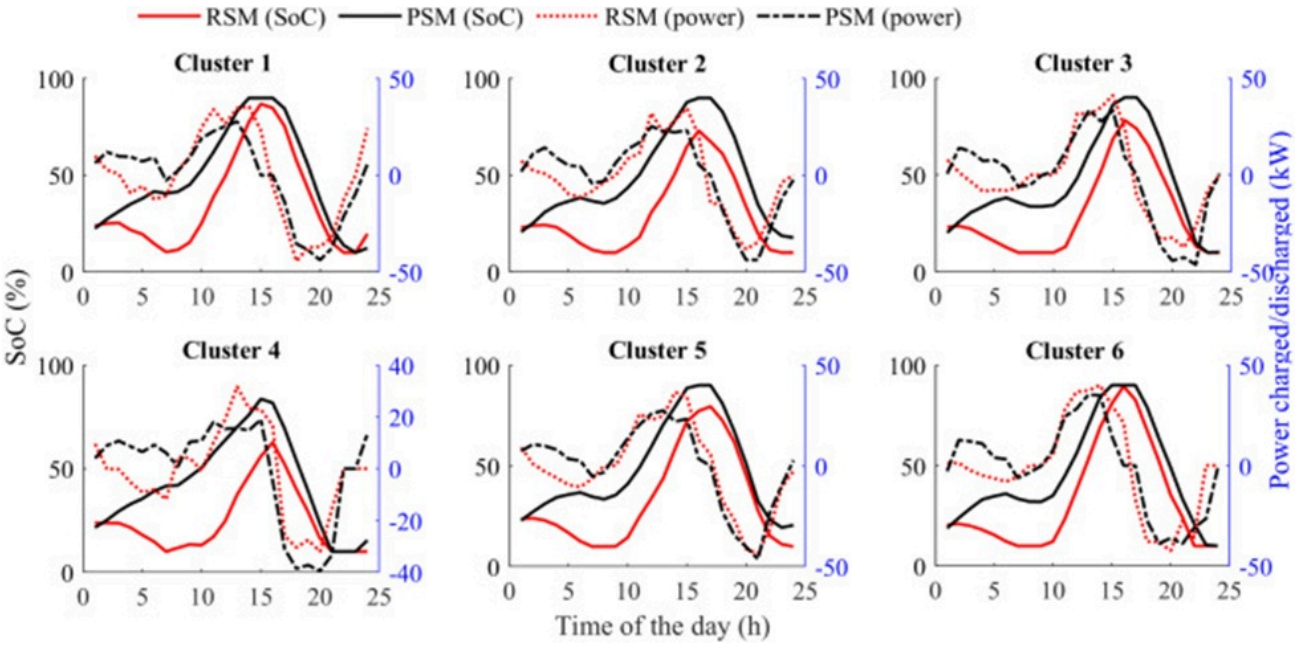


Figure 18. BESS power and SoC profiles with the integration of PV.

Conclusion

This study introduces an experience-diversity-driven deep reinforcement learning framework designed to train agents to regulate power demand for facilities equipped with solar PV and energy storage. Initial datasets are categorized into distinct clusters, accompanied by a custom selection algorithm that extracts information from every group, guaranteeing the agent interacts with a complete spectrum of data variations. Convergence evaluations reveal that this new strategy reaches stability more rapidly than techniques relying on traditional random data sampling. Furthermore, the introduced framework consistently achieved greater reward outcomes alongside improved consistency in its average reward metric. This suggests the presented methodology is highly effective at navigating its operational environment and identifying

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

optimal strategies for load regulation via battery charge and discharge management. Likewise, the new method yielded a superior flatness index across every scenario, with the difference measuring up to 32% lower when contrasted with standard practices. Across all evaluated instances, the divergence index remained at its minimum with the proposed strategy, demonstrating that the final load distribution closely mirrors the typical daily average. Relative to conventional techniques, the variation in this divergence index reached up to 15%. Comprehensive assessments confirm the exceptional overall efficacy of this new framework, stemming from its capacity to analyze representative data from the full operational setting and execute superior battery charge and discharge decisions.

Acknowledgements

This research has been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, grant number RGPIN-2017-05866.

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 22 Jan 2023 Revised: 30 Mar 2023 Accepted: 14 Apr 2023
Published online: 15 June 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Liu J, Chen X, Yang H, Li Y. Energy storage and management system design optimization for a photovoltaic integrated low-energy building. Energy. 2020;190:116424. https://doi.org/10.1016/j.energy.2019.116424.

2. United Nations Environment Programme. Global status report for buildings and construction. Tech. rep. UNEP; 2022. Available from: https://www.unep.org/resources/publication/2022-global-status-report-buildings-and-construction.

3. IEA. Electricity market report 2023. Tech. rep. IEA; 2023. Available from: https://www.iea.org/events/electricity-market-report-2023.

4. AESO. AESO net-zero emissions pathways. Tech. rep. Alberta Electric System Operator (AESO), Canada; 2022. Available from: https://www.aeso.ca/market/net-zero-emissions-pathways.

5. Hussain A, Musilek P. Resilience enhancement strategies for and through electric vehicles. Sustain Cities Soc. 2022;103788. https://doi.org/10.1016/j.scs.2022.103788.

6. Wei M, McMillan CA, de la Rue du Can S. Electrification of industry: Potential, challenges and outlook. Curr Sustain Renew Energy Rep. 2019;6:140-8. https://doi.org/10.1007/s40518-019-00136-1.

7. White PR, Rhodes JD, Wilson EJ, Webber ME. Quantifying the impact of residential space heating electrification on the Texas electric grid. Appl Energy. 2021;298:117113. https://doi.org/10.1016/j.apenergy.2021.117113.

8. Chen CY, Chai KK, Lau E. AI-assisted approach for building energy and carbon footprint modeling. Energy AI. 2021;5:100091. https://doi.org/10.1016/j.egyai.2021.100091.

9. MRC Stratistics. UAV drones – global market outlook (2016–2022). Report ID: SMRC16075; 2016.

10. Bahramara S. Robust optimization of the flexibility-constrained energy management problem for a smart home with rooftop photovoltaic and an energy storage. J Energy Storage. 2021;36:102358. https://doi.org/10.1016/j.est.2021.102358.

11. Niu J, Tian Z, Lu Y, Zhao H. Flexible dispatch of a building energy system using building thermal storage and battery energy storage. Appl Energy. 2019;243:274-87. https://doi.org/10.1016/j.apenergy.2019.03.203.

12. Liu J, Chen X, Cao S, Yang H. Overview on hybrid solar photovoltaic-electrical energy storage technologies for power supply to buildings. Energy Convers Manage. 2019;187:103-21. https://doi.org/10.1016/j.enconman.2019.02.080.

13. Terlouw T, AlSkaif T, Bauer C, van Sark W. Optimal energy management in all-electric residential energy systems with heat and electricity storage. Appl Energy. 2019;254:113580. https://doi.org/10.1016/j.apenergy.2019.113580.

14. Koskela J, Rautiainen A, Järventausta P. Using electrical energy storage in residential buildings – Sizing of battery and photovoltaic panels based on electricity cost optimization. Appl Energy. 2019;239:1175-89. https://doi.org/10.1016/j.apenergy.2019.02.015.

15. Mariano-Hernández D, Hernández-Callejo L, Zorita-Lamadrid A, Duque-Pérez O, García FS. A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. J Build Eng. 2021;33:101692. https://doi.org/10.1016/j.jobe.2020.101692.

16. Tien PW, Wei S, Darkwa J, Wood C, Calautit JK. Machine learning and deep learning methods for enhancing building energy efficiency and indoor environmental quality – a review. Energy AI. 2022;100198. https://doi.org/10.1016/j.egyai.2022.100198.

17. Golpîra H, Khan SAR. A multi-objective risk-based robust optimization approach to energy management in smart residential buildings under combined demand and supply uncertainty. Energy. 2019;170:1113-29. https://doi.org/10.1016/j.energy.2018.12.201.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

18. Javadi MS, Gough M, Lotfi M, Nezhad AE, Santos SF, Catalão JP. Optimal self-scheduling of home energy management system in the presence of photovoltaic power generation and batteries. Energy. 2020;210:118568. https://doi.org/10.1016/j.energy.2020.118568.

19. Nge CL, Ranaweera IU, Midtgård OM, Norum L. A real-time energy management system for smart grid integrated photovoltaic generation with battery storage. Renew Energy. 2019;130:774-85. https://doi.org/10.1016/j.renene.2018.06.073.

20. Lokeshgupta B, Sivasubramani S. Multi-objective home energy management with battery energy storage systems. Sustain Cities Soc. 2019;47:101458. https://doi.org/10.1016/j.scs.2019.101458.

21. Ahmad S, Naeem M, Ahmad A. Low complexity approach for energy management in residential buildings. Int Trans Electr Energy Syst. 2019;29(1):e2680. https://doi.org/10.1002/etep.2680.

22. Azuatalam D, Lee WL, de Nijs F, Liebman A. Reinforcement learning for whole-building HVAC control and demand response. Energy AI. 2020;2:100020. https://doi.org/10.1016/j.egyai.2020.100020.

23. Abedi S, Yoon SW, Kwon S. Battery energy storage control using a reinforcement learning approach with cyclic time-dependent Markov process. Int J Electr Power Energy Syst. 2022;134:107368. https://doi.org/10.1016/j.ijepes.2021.107368.

24. Alfaverh F, Denai M, Sun Y. Demand response strategy based on reinforcement learning and fuzzy reasoning for home energy management. IEEE Access. 2020;8:39310-21. https://doi.org/10.1109/ACCESS.2020.2974577.

25. Yu L, Xie W, Xie D, Zou Y, Zhang D, Sun Z, et al. Deep reinforcement learning for smart home energy management. IEEE Internet Things J. 2019;7(4):2751-62. https://doi.org/10.1109/JIOT.2019.2957289.

26. Shen R, Zhong S, Wen X, An Q, Zheng R, Li Y, et al. Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. Appl Energy. 2022;312:118724. https://doi.org/10.1016/j.apenergy.2022.118724.

27. Pinto G, Deltetto D, Capozzoli A. Data-driven district energy management with surrogate models and deep reinforcement learning. Appl Energy. 2021;304:117642. https://doi.org/10.1016/j.apenergy.2021.117642.

28. Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy. 2021;229:120725. https://doi.org/10.1016/j.energy.2021.120725.

29. Shaqour A, Hagishima A. Systematic review on deep reinforcement learning-based energy management for different building types. Energies. 2022;15(22):8663. https://doi.org/10.3390/en15228663.

30. Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy. 2020;269:115036. https://doi.org/10.1016/j.apenergy.2020.115036.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

31. Nagy Z, Henze G, Dey S, Arroyo J, Helsen L, Zhang X, et al. Ten questions concerning reinforcement learning for building energy management. Build Environ. 2023;110435. https://doi.org/10.1016/j.buildenv.2023.110435.

32. Weinberg D, Wang Q, Timoudas TO, Fischione C. A review of reinforcement learning for controlling building energy systems from a computer science perspective. Sustain Cities Soc. 2022;104351. https://doi.org/10.1016/j.scs.2022.104351.

33. Nweye K, Liu B, Stone P, Nagy Z. Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings. Energy AI. 2022;10:100202. https://doi.org/10.1016/j.egyai.2022.100202.

34. Parker-Holder J, Pacchiano A, Choromanski KM, Roberts SJ. Effective diversity in population based reinforcement learning. Adv Neural Inf Process Syst. 2020;33:18050-62.

35. Dey S, Marzullo T, Zhang X, Henze G. Reinforcement learning building control approach harnessing imitation learning. Energy AI. 2023;14:100255. https://doi.org/10.1016/j.egyai.2023.100255.

36. Hong ZW, Shann TY, Su SY, Chang YH, Fu TJ, Lee CY. Diversity-driven exploration strategy for deep reinforcement learning. Adv Neural Inf Process Syst. 2018;31.

37. Pereira T, Abbasi M, Ribeiro B, Arrais JP. Diversity oriented deep reinforcement learning for targeted molecule generation. J Cheminformatics. 2021;13:1-17. https://doi.org/10.1186/s13321-021-00533-7.

38. Dai T, Du Y, Fang M, Bharath AA. Diversity-augmented intrinsic motivation for deep reinforcement learning. Neurocomputing. 2022;468:396-406. https://doi.org/10.1016/j.neucom.2021.10.055.

39. Kanishev K. PyIDF: Diversity of experiences in reinforcement learning. 2019. Available from: https://medium.com/imandra/pyidf-diversity-of-experiences-in-reinforcement-learning-8d59f60f59ed.

40. Nazari M, Hussain A, Musilek P. Applications of clustering methods for different aspects of electric vehicles. Electronics. 2023;12(4):790. https://doi.org/10.3390/electronics12040790.

41. Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmos Pollut Res. 2020;11(1):40-56. https://doi.org/10.1016/j.apr.2019.09.009.

42. Hussain A, Bui VH, Kim HM. Deep reinforcement learning-based operation of fast charging stations coupled with energy storage system. Electr Power Syst Res. 2022;210:108087. https://doi.org/10.1016/j.epsr.2022.108087.

43. Bui VH, Su W. Real-time operation of distribution network: A deep reinforcement learning-based reconfiguration approach. Sustain Energy Technol Assess. 2022;50:101841. https://doi.org/10.1016/j.seta.2021.101841.

44. Bui VH, Hussain A, Su W. A dynamic internal trading price strategy for networked microgrids: A deep reinforcement learning-based game-theoretic approach. IEEE Trans Smart Grid. 2022;13(5):3408-21. https://doi.org/10.1109/TSG.2022.3167743.

45. Hussain A, Musilek P. Utility-scale energy storage system for load management under high penetration of electric vehicles: A marginal capacity value-based sizing approach. J Energy Storage. 2022;56:105922.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press

https://doi.org/10.1016/j.est.2022.105922.

46. Taheri S, Jooshaki M, Moeini-Aghtaie M. 8 years of hourly heat and electricity demand for a residential building. IEEE Dataport. 2021.
https://doi.org/10.21227/dfvb-re49.

47. Renewables Ninja. Available from: https://www.renewables.ninja.

48. Elbow Method for optimal value of k in KMeans. GeeksforGeeks. Available from: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.

49. Teoh TT, Rong Z. Clustering. In: Artificial intelligence with Python. Singapore: Springer Singapore; 2022. p. 213-8.
https://doi.org/10.1007/978-981-16-8615-3_9.

50. ML – T-distributed Stochastic Neighbor Embedding (t-SNE) Algorithm. GeeksforGeeks. Available from: https://www.geeksforgeeks.org/ml-t-distributed-stochastic-neighbor-embedding-t-sne-algorithm/.

51. Viola LG. Clustering electricity usage profiles with K-means. Towards Data Science. Available from: https://towardsdatascience.com/clustering-electricity-profiles-with-k-means-42d6d0644d00.

© 2024 The Author(s). | Published by Institute of Smart Energy Systems Press